JOURNAL OF MULTIVARIATE ANALYSIS **58**, 133–150 (1996)
ARTICLE NO. 0043

Incorporating Extra Information in Nonparametric Smoothing

Rong Chen

Department of Statistics, Texas A&M University, College Station, Texas 77843

Nonparametric estimation of conditional mean functions has been studied extensively in the literature. This paper addresses the question of how to use extra informations to improve the estimation. Particularly, we consider the situation that the conditional mean function $E(Z \mid X)$ is of interest and there is an auxiliary variable available which is correlated with both X and Z. A two-stage kernel smoother is

when using the Nadaraya–Watson estimator directly without the auxiliary variable. A simulation study is also carried out to illustrate the procedure. © 1996 Academic Press, Inc.

1. INTRODUCTION

There have been many studies on nonparametric estimation of the conditional mean functions using Kernel smoothing, spline smoothing and local polynomial methods. For example, see Eubank [4], Härdle [6], Collomb [3], Härdle and Marron [7], Hall [8], Cleveland and Devlin [2], Fan [5], Stone [13, 14] etc. and the references therein. However, the question of how to incorporate extra information into estimation has not been investigated. In this paper we consider the case where the conditional mean function $E(Z \mid X)$ is of interest and there is an auxiliary variable Y available which is correlated with both X and Z. We propose a two-stage kernel smoother to improve the estimation by incorporating the auxiliary variable Y.

This problem is of interest in many situations. For example, in measurement error models we observe three variables, the response variable Z, a true predictor Y and a predictor X measured with error. Since the true

Received October 19, 1994; revised February 1996.

AMS 1991 subject classifications: primary 62G07; secondary 62H12.

Key words and phrases: kernel smoothing, mean squared errors, two-stage smoother, auxiliary variable.

0047-259X/96 $18.00

predictor usually is more expensive and difficult to measure in practice, we are interested in predicting Z given only the less expensive predictor X. This can be done by using the least square predictor $E(Z \mid X)$. Note that, if we do not observe the Y variable, the conditional expectation $E(Z \mid X)$ can be estimated nonparametrically by directly smoothing Z on X. In our problem setting, we do observe Y and we wish to utilize the extra information to estimate $E(Z \mid X)$ more accurately.

We will show that the proposed two-stage kernel smoother which incorporates the information in Y has smaller pointwise and integrated asymptotically mean squared error than the corresponding kernel estimator that does not use this information. The amount of improvement depends on two things: the ratio of the conditional variances $E(\mathrm{Var}(Z \mid X, Y))$ and $\mathrm{Var}(E(Z \mid X, Y))$ and a function of the partial derivatives of the function $E(Z \mid X, Y)$ with respect to x.

The following observation motivates the proposed procedure. Let $f(x, y) = E(Z \mid X = x, Y = y)$, $Z^* = f(X, Y)$ and $Z_i^* = f(X_i, Y_i)$. We have

$$m(x) \equiv E(Z \mid X = x) = E(E(Z \mid X, Y) \mid X = x)$$
$$= E(f(X, Y) \mid X = x) = E(Z^* \mid X = x).$$

Since $\mathrm{Var}(Z^* \mid X = x) \leqslant \mathrm{Var}(Z \mid X = x)$, it is obvious that smoothing with the pairs (Z_i^*, X_i), $i = 1, ..., n$, would provide a more accurate estimator than using the pairs (Z_i, X_i). But, most of the time the function $f(\cdot, \cdot)$ and Z^* are unknown and unobservable. However, note that the difference of Z and Z^* is of order $O_p(1)$. Hence, if we estimate the function $f(\cdot, \cdot)$ with a suitable estimator $\hat{f}(\cdot, \cdot)$ that has a smaller error rate (with some bias), then we can use the pairs $(\hat{Z}_i^*, X_i)$ to estimate $m(\cdot)$ where $\hat{Z}_i^* = \hat{f}(X_i, Y_i)$. In this way, smaller errors may be achieved. We shall prove that this is indeed the case.

The above observation motivates the following estimator of $E(Z \mid X)$, which will be referred to as "two-stage smoother". It is defined as

$$\hat{m}_{h_1, h_2, h_3}(x) = \frac{\sum_{k=1}^n K((x - X_k)/h_3)\, \hat{Z}_{k, h_1, h_2}^*}{\sum_{k=1}^n K((x - X_k)/h_3)}, \tag{1}$$

where $\hat{Z}_{k, h_1, h_2}^* = \hat{f}_{h_1, h_2}(X_k, Y_k)$ and

$$\hat{f}_{h_1, h_2}(x, y) = \frac{\sum_{j=1}^n K((x - X_j)/h_1)\, K((y - Y_j)/h_2)\, Z_j}{\sum_{j=1}^n K((x - X_j)/h_1)\, K((y - Y_j)/h_2)}.$$

Note that $\hat{f}$ is a regular two-dimensional Nadaraya–Watson (N-W) estimator (Nadaraya [9], Watson [15]) of $f(x, y) = E(Z \mid X = x, Y = y)$.

We shall compare the proposed estimator (1) with the regular N-W estimator

$$\tilde{m}_h(x) = \frac{\sum_{k=1}^{n} K((x - X_k)/h)\, Z_k}{\sum_{k=1}^{n} K((x - X_k)/h)}, \tag{2}$$

which does not utilize the information in Y variable.

The rest of the paper is organized as follows. In Section 2, the asymptotic pointwise and integrated mean squared error of the two-stage estimator (1) are compared to that of the N-W estimator (2). Empirical comparisons via a small simulation study are described in Section 3 and a brief summary is presented in Section 4. Conditions and proof of the theorems are collected in Section 5.

2. ASYMPTOTIC PROPERTIES OF THE TWO-STAGE SMOOTHER

Throughout the paper, the following notations are used. First, $p(x, y)$ is used to denote the joint density of (X, Y) and $p(x)$ the marginal density of X. The conditional means and variances are defined as follows:

$$m(x) = E(Z \mid X = x), \qquad f(x, y) = E(Z \mid X = x,\ Y = y),$$

$$v(x, y) = \mathrm{Var}(Z \mid X = x,\ Y = y), \qquad u(x) = \mathrm{Var}(f(X, Y) \mid X = x),$$

$$w(x) = E(v(X, Y) \mid X = x), \qquad \text{and} \quad \sigma^2(x) = \mathrm{Var}(Z \mid X = x).$$

It is important to note that $\sigma^2(x) = u(x) + w(x)$.

Let $K(\cdot)$ be a bounded kernel function with finite support and $\int K(z)\, dz = 1$. To simplify our notation, define constants $k_1 = \int K^2(z)\, dz$, $k_2 = \int z^2 K(z)\, dz$. In addition, the following three functions are important ingredients of our results.

$$s_1(x) = m''(x)\, p(x) + 2m'(x)\, p'(x),$$

$$s_2(x) = \int t_1(x, y)\, dy, \qquad \text{and} \qquad d(\theta) = \frac{1}{k_1} \int L_\theta^2(z)\, dz,$$

where

$$t_1(x, y) = \frac{\partial f^2(x, y)}{\partial x^2}\, p(x, y) + 2\, \frac{\partial f(x, y)}{\partial x}\, \frac{\partial p(x, y)}{\partial x}, \tag{3}$$

and

$$L_\theta(z) = \int K(z + \theta u)\, K(u)\, du. \tag{4}$$

THEOREM 2.1. *Under conditions* (C1)–(C5) *presented in Section* 5, *if* $h_3 \to 0$, $nh_3 \to \infty$, $h_1 = \theta h_3$, $h_2 = o(h_3)$, *and* $nh_1 h_2 \to \infty$, *then for a fixed* $x \in \mathscr{A}$, *the asymptotic pointwise mean squared error of estimator* (1) *is*

$$E(\hat{m}(x) - m(x))^2 = \frac{D_1(x, \theta)}{nh_3} + \frac{1}{4} D_2(x, \theta)\, h_3^4 + o\left(\frac{1}{nh_3} + h_3^4\right), \tag{5}$$

where

$$D_1(x, \theta) = k_1(u(x) + d(\theta)\, w(x))/p(x)$$

and

$$D_2(x, \theta) = k_2^2(s_1(x) + s_2(x)\, \theta^2)^2/p^2(x).$$

It is well known (e.g. Collomb [3], Stone [14], Härdle [6]) that the direct smoother (2) have the property

$$E(\tilde{m}(x) - m(x))^2 = \frac{D_3(x)}{nh} + \frac{1}{4} D_4(x)\, h^4 + o\left(\frac{1}{nh} + h^4\right),$$

where

$$D_3(x) = k_1 \sigma^2(x)/p(x) = k_1(u(x) + w(x))/p(x)$$

and

$$D_4(x) = k_2^2 s_1^2(x)/p^2(x).$$

Comparing $D_1(x)$ and $D_3(x)$, we see that the two-stage smoothing reduces the asymptotic variances, since $d(\theta) < 1$. There is an extra term $s_2(x)\theta^2$ in the asymptotic bias term $D_2(x)$. Its effect depends on the sign of $s_2(x)$, comparing to $s_1(x)$. We will discuss it in detail later in the remarks.

Let $r(x)$ be the ratio of minimum asymptotic pointwise mean squared errors of estimators (1) and (2) and θ^* be the minimizer of

$$(u(x) + w(x)\, d(\theta))^4\, (s_1(x) + s_2(x)\, \theta^2)^2 \tag{6}$$

with respect to θ.

THEOREM 2.2. *Under the conditions of Theorem* 2.1,

(i) *if* $s_1(x)\, s_2(x) > 0$, *then*

$$r(x) = \left(1 - \frac{1 - d(\theta^*)}{1 + (u(x)/w(x))}\right)^{4/5} \left(1 + \frac{s_2(x)}{s_1(x)}\, \theta^{*2}\right)^{2/5}. \tag{7}$$

(ii) *If* $s_1(x)\, s_2(x) < 0$, *then the minimum asymptotic pointwise mean squared errors of the estimator* (1) *has smaller order than that of the estimator* (2).

(iii) *If* $s_1(x) = 0$, *then* $r(x) = 1$ *with* $\theta = 0$.

(iv) *If* $s_1(x) \neq 0$ *and* $s_2(x) = 0$, *then* $r(x) = (u(x)/\sigma^2(x))^{4/5}$ *with* $\theta = \infty$.

Remarks. 1. The function $d(\theta)$ plays an important role in our results. Figure 1 shows the function $d(\cdot)$ with respect to Uniform, Triangle, Epanechnikov and Quartic kernels. Note that they are essentially the same. The normalizing constant k_1 is included in the definition of $d(\theta)$ so that $d(0) = 1$. It is easy to show that, for a fixed θ, the function L_θ defined in (4) behaves exactly as a kernel function, with $\int L_\theta(z)\, dz = 1$ and $\int z L_\theta(z)\, dz = 0$. If the kernel K has bounded support, then L_θ has bounded support as well. In addition, it can be easily shown that $d(\theta)$ goes to zero

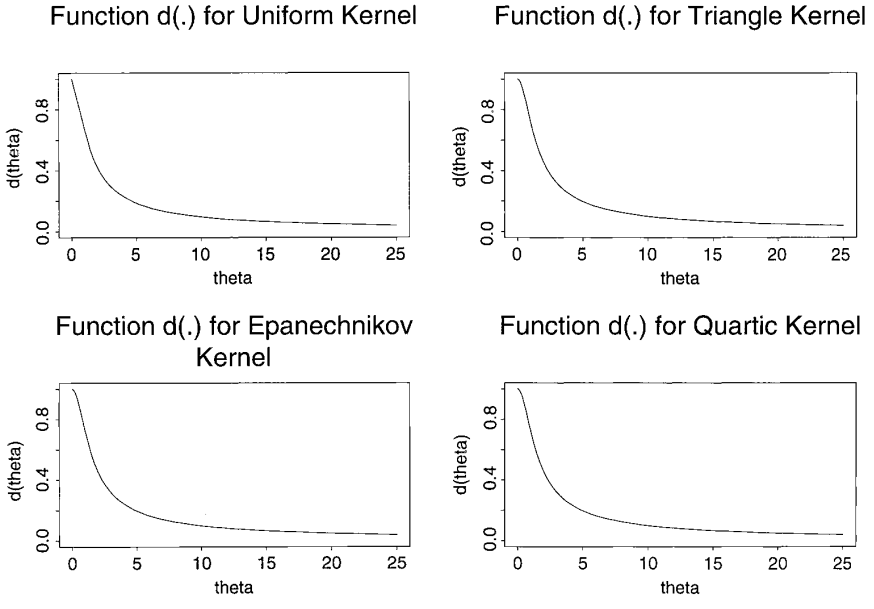

FIG. 1. The function $d(\theta)$ with respect to uniform, triangle, epanechnikov, and quartic kernels.

as θ goes to infinity. When θ increases, $d(\theta)$ decreases for many commonly used kernel functions. But it is not so in general.

2. Figure 2 shows that the contour plot of $r(\cdot)$ in (7) as the function of the ratios $u(x)/w(x)$ and positive $s_2(x)/s_1(x)$ for Uniform kernel. The function is essentially the same for other kernels, due to the similarity of $d(\theta)$. From the theorem, we can see that, when $s_1(x)\, s_2(x) > 0$, the mean squared errors of the two smoothers have the same order, but that of $\hat{m}(x)$ is proportionately smaller than that of $m(x)$. The amount of improvement depends on the ratios $u(x)/w(x)$ and $s_2(x)/s_1(x)$.

3. When $s_1(x) = 0$, the bias of the direct smoother (2) is of higher order of h_3^4. On the other hand, any smoothing using the Y variable will create a bias of order h_3^4. Hence, it is the best not to use the two-stage smoother, i.e. setting both $h_1 = h_2 = 0$.

4. If X and Z are conditionally independent given Y, then $f(x, y) = f(y)$, which implies $s_2(x) = 0$. In this case, the triple (X, Y, Z) has

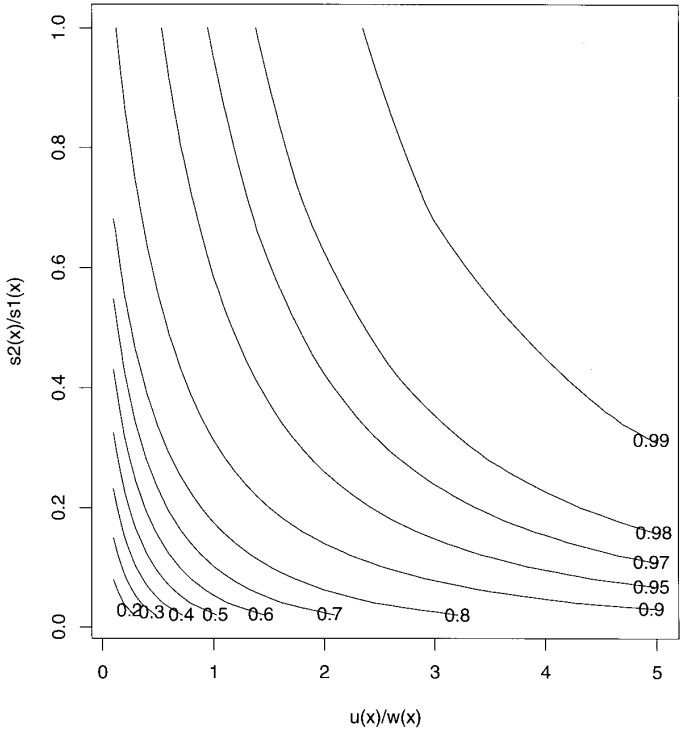

FIG. 2. The contour plot of $r(\cdot)$ (the improvement rate) in (7) as the function of the ratios $u(x)/w(x)$ and positive $s_2(x)/s_1(x)$ for uniform kernel.

the Markov property $X \to Y \to Z$. Since $s_2(x) = 0$, the θ that minimizes (6) is ∞, i.e. $h_1 = \infty$. This is saying that, only one-dimensional smoothing on (Z, Y) should be carried out in the first stage smoothing. Since $d(\infty) = 0$, the conditional variance $w(x) = \mathrm{Var}(E(Z \mid X, Y))$ disappears in the expression of D_1. The result is equivalent to smoothing Z^* on X, with *known* function $f(x, y)$. This actually is the result of Chen [1] where a similar multi-stage smoother is introduced for a multi-step prediction problem in a Markovian structure.

5. The greatest improvement can be made when $s_1(x)$ and $s_2(x)$ are of opposite signs. In this case, not only the variance is reduced, but also the biases created in the two stages of smoothing tends to cancel each other out. Theoretically, an order lower than $n^{-4/5}$ is achievable. This is impossible in practice since one must have the exact knowledge of $s_1(x)/s_2(x)$ in order to obtain the correct θ. Nevertheless, with θ^2 close to $-s_1(x)/s_2(x)$, the improvement can be significant. Also note that if the integrated mean squared error is of concern, an order smaller than $n^{-4/5}$ cannot be achieved, unless $s_1(x)/s_2(x)$ is a negative constant over the range of interest.

6. If $w(x) = 0$ and $s_1(x)$ and $s_2(x)$ are of the same sign, then the θ that minimizes (6) is $\theta = 0$. This is saying that, if $Z = f(X, Y)$, one should not do the first stage smoothing at all since there is no variance to reduce while smoothing only creates extra bias. When $s_1(x)$ and $s_2(x)$ are of opposite signs, the two-stage smoothing is still beneficial since the bias can be reduced. It actually can be used as a bias reduction tool.

7. Under mild conditions, we have

$$s_1(x) = s_2(x) + \int f(x, y) \frac{\partial p^2(x, y)}{\partial x^2}\, dy - m(x)\, p''(x).$$

Hence, when X and Y are independent, we have $s_1(x) = s_2(x)$. In this case, the maximum improvement can be shown to be $r = 0.945$ in (7) for the Uniform kernel.

The next theorems compare the integrated mean squared errors of the proposed two-stage smoother (1) and the N-W estimator (2).

THEOREM 2.3. *Under the conditions of Theorem* 2.1, *the asymptotic integrated mean squared error of estimator* (1) *is*

$$\int_{\mathscr{A}} E(\hat{m}(x) - m(x))^2\, dx = \frac{D_1(\theta)}{nh_3} + \frac{1}{4} h_3^4 D_2(\theta)\, n^{-4/5} + o\left(\frac{1}{nh_3} + h_3^4\right), \quad (8)$$

where $D_i(\theta) = \int_{\mathscr{A}} D_i(x, \theta)\, dx$. *The asymptotic optimal bandwidth is* $h_3 = (D_1(\theta^*)/D_2(\theta^*))^{1/5} n^{-1/5}$ *where* θ^* *minimizes*

$$\left(\int_{\mathscr{A}} \frac{u(x) + w(x)\, d(\theta)}{p(x)}\, dx \right)^4 \left(\int_{\mathscr{A}} \left(\frac{s_1(x) + s_2(x)\theta^2}{p(x)} \right)^2 dx \right),$$

with respect to θ. *The corresponding asymptotically optimal integrated mean squared error is*

$$\int_{\mathscr{A}} E(\hat{m}(x) - m(x))^2\, dx = 1.25 D_1(\theta^*)^{4/5}\, D_2(\theta^*)^{1/5}\, n^{-4/5} + o(n^{-4/5}).$$

THEOREM 2.4. *Under the conditions of Theorem* 2.3, *the ratio of minimum asymptotic integrated mean squared errors of estimators* (1) *and* (2) *is*

$$r = \left(1 - \frac{1 - d(\theta^*)}{1 + (u/w)} \right)^{4/5} \left(1 + 2\frac{s_{12}}{s_{11}}\theta^{*2} + \frac{s_{22}}{s_{11}}\theta^{*4} \right)^{1/5}, \tag{9}$$

where $u = \int_{\mathscr{A}} u(x)/p(x)\, dx$, $w = \int_{\mathscr{A}} w(x)/p(x)\, dx$, $s_{kl} = \int_{\mathscr{A}} s_k(x)\, s_l(x)/p^2(x)\, dx$ *and the* θ^* *is that in Theorem* 2.3.

3. SIMULATION STUDY

In this section, a simulation study is carried out to compare estimators (1) and (2). Using the following model:

$$X \sim \text{Uniform}(-0.5, 0.5), \qquad Y = 10X + \varepsilon, \qquad \text{and} \qquad Z = 3\sin(Y)\cos(X) + e,$$

where $\varepsilon \sim U(-0.5, 0.5)$ and $e \sim U(-1.5, 1.5)$. Two hundred samples, each with six hundred observations are generated. For each sample, estimators (1) and (2) are evaluated at 160 equally spaced grids on the interval $(-0.4, 0.4)$ using Triangle kernels. Leave-one-out cross validation is used to choose the bandwidth for both estimators. For each of the 160 points, the squared errors for estimating the true conditional mean $m(x) = 6\sin(0.5)\sin(10x)\cos(x)$ are computed and averaged across the 160 points. Denote these averages by r_1 and r_2, for the two-stage estimator (1) and the ordinary N-W estimator (2) respectively. Table I shows that the percentiles for the ratios r_1/r_2 from the two hundred samples. The theoretical improvement of the integrated mean squared error using (9) for the above model

on the interval $(-0.4, 0.4)$ is $r = 0.139$ for $\theta^* = 3.2$. The leave-one-out cross-validation shows an average of 6.5 for $\theta = h_1/h_3$ in our simulation.

The computation of the leave-one-out cross validation criterion for the two-stage estimator is very intensive. In Table I we also show the results of an alternative approach where ordinary cross-validation computations are used for each stage of the smoothing separately. First, the optimal bandwidth (h_1^*, h_2^*) for the first stage two-dimensional smoothing is found using ordinary leave-one-out cross validation. Then the function $\hat{f}(\cdot, \cdot)$ is estimated using an adjusted bandwidth $(c_1 h_1^*, c_2 h_2^*)$ for some constants c_1 and c_2. The adjustment is needed since the optimal bandwidth for the first stage smoothing may not be optimal over all. After the first stage smoothing, the bandwidth h_3 for the second stage smoothing is then chosen to be the optimal cross validation bandwidth for the pair $(\hat{Z}_i^*, X_i)$, where $\hat{Z}_i^* = \hat{f}(X_i, Y_i)$. In Table I, we show the simulation results for some of the combinations of (c_1, c_2).

From the table we can see that the bandwidth selected by leave-one-out cross validation does well. Over 65% of the time the proposed two-stage estimator improves the mean squared error, with a median of 10% improvement. It also shows that with small c_2, we can actually obtain reasonable results using cross-validation criterion separately for each stage of smoothing. This saves computation time. We have tried several other examples and observed similar results.

TABLE I

Percentile:	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
CV:	0.66	0.75	0.80	0.83	0.90	0.94	1.04	1.09	1.25
(c_1, c_2)									
0.5, 0.1)	0.68	0.73	0.77	0.81	0.86	0.91	0.97	1.09	1.25
(1.0, 0.1)	0.67	0.72	0.76	0.81	0.86	0.90	0.97	1.09	1.25
(2.0, 0.1)	0.67	0.71	0.76	0.81	0.86	0.90	0.97	1.09	1.25
(0.5, 0.2)	0.64	0.71	0.77	0.81	0.85	0.90	1.01	1.12	1.33
(1.0, 0.2)	0.63	0.71	0.76	0.81	0.85	0.90	1.01	1.12	1.35
(2.0, 0.2)	0.63	0.70	0.75	0.80	0.86	0.90	1.01	1.12	1.36
(0.5, 0.5)	0.68	0.74	0.81	0.85	0.91	0.95	1.04	1.18	1.37
(1.0, 0.5)	0.67	0.75	0.82	0.86	0.91	0.95	1.03	1.18	1.38
(2.0, 0.5)	0.67	0.75	0.82	0.86	0.91	0.95	1.03	1.18	1.38
(0.5, 1.0)	0.90	1.00	1.06	1.12	1.20	1.27	1.34	1.47	1.73
(1.0, 1.0)	0.91	1.02	1.09	1.16	1.23	1.30	1.37	1.51	1.78
(2.0, 1.0)	0.91	1.03	1.11	1.18	1.25	1.32	1.39	1.54	1.83

4. SUMMARY

In this paper, we proposed a two-stage kernel smoother to estimate the conditional mean function when information on an auxiliary variable is available. It is shown, both theoretically and empirically, that the proposed smoother has smaller asymptotic mean squared error than the N-W estimator.

There is one interesting observation. If multiplicative kernel density estimators are used to estimate $p_X(x)$, $p_{X,Y}(x, y)$ and $p_{X,Y,Z}(x, y, z)$, then an analogous derivation to that of the N-W estimator as in Eubank [4, pages 169–170] results

$$m(x) \sim \frac{\sum_{k=1}^{n} K((x-X_k)/h_1)\, \tilde{Z}_k}{\sum_{k=1}^{n} K((x-X_k)/h_1)},$$

where

$$\tilde{Z}_k = \frac{\sum_{j=1}^{n} K((x-X_j)/h_1)\, K((Y_k-Y_j)/h_2)\, Z_j}{\sum_{i=1}^{n} K((x-X_i)/h_1)\, K((Y_k-Y_i)/h_2)}.$$

This estimator differs from (1) in that here $\tilde{Z}_k = \hat{f}(x, Y_k)$, instead of $\hat{Z}_k = \hat{f}(X_k, Y_k)$. However, it can be proved that this estimator does not lead to an improvement in asymptotic mean squared error. This means that the information in the auxiliary variable must be used in the right way in order to improve the mean squared error.

We also note that the kernel estimator used in our approach can be replaced with local polynomial estimators. Although details will be different, the effect of the two-stage smoothing remains the same.

5. ASSUMPTIONS AND PROOFS

In addition to all the functions defined in Section 2, the following functions are needed for technical reasons:

$$t_2(x, y) = \frac{\partial f^2(x, y)}{\partial y^2} p(x, y) + 2 \frac{\partial f(x, y)}{\partial y} \frac{\partial p(x, y)}{\partial y}, \tag{10}$$

$$g_1(x) = E_Y\left[\frac{v(X, Y)}{p^2(X, Y)} \,\middle|\, X = x \right], \qquad g_2(x) = E_Y\left[\frac{v(X, Y)}{p(X, Y)} \,\middle|\, X = x \right],$$

$$g_3(x) = E_Y\left[\frac{1}{p(X, Y)} \,\middle|\, X = x \right], \qquad \text{and} \qquad g_4(x) = \int t_2(x, y)\, dy.$$

For the purpose of estimating $m(\cdot)$, the set of interest is assumed to be a finite interval $\mathscr{A}$, i.e., we are only interested in estimating $E(Z \mid X = x)$ for $x \in \mathscr{A}$. Let $\mathscr{A}_\varepsilon$ be the set of all the points in $\mathfrak{R}$ which are distinct less than ε from $\mathscr{A}$. The following conditions are assumed.

(C1) $K(\cdot)$ is a bounded density function with compact support satisfying $\int zK(z)\,dz = 0$ for $k_1, k_2 < \infty$.

(C2) The marginal density $p(x)$ of X has finite second derivative and is bounded away from zero in $\mathscr{A}_\varepsilon$. The joint density $p(x, y)$ of (X, Y) has finite support $\mathscr{B}_x$ in the Y variable. Let $\mathscr{B}_\varepsilon = \bigcup_{x \in \mathscr{A}_\varepsilon} \mathscr{B}_x$. The joint density of $p(x, y)$ is twice differentiable with finite second partial derivatives both in the x and y variables.

(C3) The function $m(\cdot)$ is twice differentiable and the second derivative is Hölder continuous such that $|m''(x_1) - m''(x_2)| \leqslant c_1 |x_1 - x_2|^{\gamma_1}$ in $\mathscr{A}_\varepsilon$ for $\gamma_1 > 0$. The function $f(x, y)$ is twice differentiable in both the x and the y variables and the second derivatives are Hölder continuous such that $|(\partial f^2(x, y_1)/\partial y^2) - (\partial f^2(x, y_2)/\partial y^2)| \leqslant c_2 |y_1 - y_2|^{\gamma_2}$, uniformly for $x \in \mathscr{A}_\varepsilon$, and $|(\partial f^2(x_1, y)/\partial x^2) - (\partial f^2(x_2, y)/\partial x^2)| \leqslant c_3 |x_1 - x_2|^{\gamma_3}$, uniformly for $y \in \mathscr{B}_\varepsilon$ with $\gamma_2 > 0$ and $\gamma_3 > 0$.

(C4) The functions $u(x)$, $w(x)$, $s_1(x)$ and $s_2(x)$ are all well defined and Hölder continuous such that $|u(x_1) - u(x_2)| < c_4 |x_1 - x_2|^{\gamma_4}$, $|w(x_1) - w(x_2)| < c_5 |x_1 - x_2|^{\gamma_5}$, $|s_1(x_1) - s_1(x_2)| < c_6 |x_1 - x_2|^{\gamma_6}$, $|s_2(x_1) - s_2(x_2)| < c_7 |x_1 - x_2|^{\gamma_7}$ for $\gamma_i > 0$, $i = 4, ..., 7$. The function $v(x, y)$ is Hölder continuous in both x and y variables such that $|v(x_1, y) - v(x_2, y)| < c_8 |x_1 - x_2|^{\gamma_8}$ and $|v(x, y_1) - v(x, y_2)| < c_9 |y_1 - y_2|^{\gamma_9}$ for $\gamma_8 > 0$ and $\gamma_9 > 0$ uniformly in $x \in \mathscr{A}_\varepsilon$ and $y \in \mathscr{B}_\varepsilon$.

(C5). The functions $g_i(x)$, $i = 1, ..., 4$ are well defined and bounded in $\mathscr{A}_\varepsilon$.

We will adopt the conventional notation $K_h(u) = h^{-1}K(u/h)$. Let

$$\hat{p}(x) = n^{-1} \sum_{k=1}^{n} K_h(x - X_k),$$

$$\text{and} \quad \hat{p}(x, y) = n^{-1} \sum_{k=1}^{n} K_{h_1}(x - X_k)\, K_{h_{23}}(y - Y_k).$$

LEMMA 5.1. *Under conditions* C1–C5, $h \to 0$, $nh \to \infty$, $h_1 \to 0$, $h_2 \to 0$ *and* $nh_1h_2 \to \infty$, *we have*

$$|\hat{p}(x) - p(x)| \to 0 \qquad \textit{and} \qquad |\hat{p}(x, y) - p(x, y)| \to 0 \textit{ in probability.}$$

Conditional on X_1, Y_1, *we have*

$$|\hat{p}(X_1, Y_1) - p(X_1, Y_1)| \to 0 \text{ in probability.}$$

These are well known results. For example, see Parzen [10], Silverman [12] or Scott [11].

In what follows, $A_n \sim B_n$ means $A_n = B_n + o_p(B_n)$, i.e. A_n equals B_n plus a term that goes to zero in probability faster than B_n as n goes to ∞. Note that if $A_n \sim B_n$ and B_n has finite support, then $E(A_n) = E(B_n) + o(E(B_n))$. In this case, we write $E(A_n) \sim E(B_n)$ as well. Also note that if $A_n \sim B_n$, then $A_n^2 \sim B_n^2$.

LEMMA 5.2. *Under conditions* C1–C5, $h \to 0$ *and* $nh \to \infty$, *then for any identically distributed random variables* W_i *we have*

$$\frac{\sum_{i=1}^n K_h(x - X_i)\, W_i}{\sum_{i=1}^n K_h(x - X_i)} \sim \frac{1}{n} \frac{\sum_{i=1}^n K_h(x - X_i)\, W_i}{p(x)}.$$

Proof. Following Härdle and Marron [7],

$$\frac{\sum_{i=1}^n K_h(x - {}_i)\, W_i}{\sum_{i=1}^n K_h(x - X_i)} = \frac{1}{n} \frac{\sum_{i=1}^n K_h(x - X_i)\, W_i}{p(x)} + \frac{1}{n} \frac{\sum_{i=1}^n K_h(x - X_i)\, W_i}{p(x)} \left(\frac{p(x)}{\hat{p}(x)} - 1 \right).$$

By Lemma 5.1, it is easy to see that the second term is negligible compared to the first.

LEMMA 5.3. *Under conditions* C1–C5, $h_1 \to 0$, $h_2 \to 0$ *and* $nh_1h_2 \to \infty$, *for any identically distributed random variables* Z_j, *we have*

$$\frac{\sum_{j=1}^n K_{h_1}(X_1 - X_j)\, K_{h_2}(Y_1 - Y_j)\, Z_j}{\sum_{j=1}^n K_{h_1}(X_1 - X_j)\, K_{h_2}(Y_1 - Y_j)} \sim \frac{1}{n} \frac{\sum_{j=1}^n K_{h_1}(X_1 - X_j)\, K_{h_2}(Y_1 - Y_j)\, Z_j}{p(X_1, Y_1)}.$$

LEMMA 5.4. *Under condition* (C1)–(C3), *and* $h_1 = o(1)$, $h_2 = o(1)$, *we have*

$$K(z_1)\, K(z_2)(f(X - h_1 z_1, Y - h_2 z_2) - f(X, Y)) = K(z_1)\, K(z_2)\, o_p(1),$$

and

$$K(z_1)\, K(z_2)\, p(X - h_1 z_1, Y - h_2 z_2) = K(z_1)\, K(z_2)(p(X, Y) + o_p(1)),$$

This comes from conditions (C3) and (C2) and the fact that $K(\cdot)$ has finite support.

LEMMA 5.5. *Under conditions* (C1)–(C5) *and* $h_1 = o(1)$, $h_2 = o(1)$ *we have*

$$\begin{aligned}\int K(z_1)\,K(z_2)&(f(X_1 - h_1 z_1,\, Y_1 - h_2 z_2)\\ &-f(X_1,\, Y_1))\,p(X_1 - h_1 z_1,\, Y_1 - h_2 z_2)\,dz_1\,dz_2\\ &= \tfrac{1}{2}k_2 h_1^2 t_1(X_1,\, Y_1) + \tfrac{1}{2}k_2 h_2^2 t_2(X_1,\, Y_1) + o_p(h_1^2 + h_2^2 + h_1 h_2),\end{aligned}$$

where $t_1(\cdot,\cdot)$, $t_2(\cdot,\cdot)$ *are defined in* (3) *and* (10).

Since $K(\cdot)$ has finite support, we can treat z_1 and z_2 as bounded. Then, using a Taylor expansion and the fact that $\int zK(z)\,dz = 0$ and $\int z^2 K(z)\,dz = k_2$, the lemma can be easily derived.

Proof of Theorem 2.1. Let $\varepsilon_j = Z_j - f(X_j, Y_j)$. We have ε_i i.i.d with $E(\varepsilon_j \mid X_j, Y_j) = 0$ and $\mathrm{Var}(\varepsilon_j \mid X_j, Y_j) = v(X_j, Y_j)$. In the following derivation, we will repeatedly use Lemma 5.2 to 5.5 and the facts that (X_i, Y_i, Z_i) are i.i.d and $\int zK(z)\,dz = 0$.

Let $Z_k^* = f(X_k, Y_k)$ and $\hat{Z}_k^* = \hat{f}(X_k, Y_k)$. For a fixed $x \in \mathscr{A}$,

$$\begin{aligned}&E[\hat{m}(x) - m(x)]^2\\ &\quad= E\left[\left(\hat{m}(x) - \frac{\sum_{k=1}^n K_{h_3}(x - X_k)\,Z_k^*}{\sum_{k=1}^n K_{h_3}(x - X_k)}\right)\right.\\ &\qquad\left.+\left(\frac{\sum_{k=1}^n K_{h_3}(x - X_k)\,Z_k^*}{\sum_{k=1}^n K_{h_3}(x - X_k)} - m(x)\right)\right]^2\\ &\quad= E\left[\frac{\sum_{k=1}^n K_{h_3}(x - X_k)(\hat{Z}_k^* - Z_k^*)}{\sum_{k=1}^n K_{h_3}(x - X_k)}\right]^2\\ &\qquad+ E\left[\frac{\sum_{k=1}^n K_{h_3}(x - X_k)(Z_k^* - m(x))}{\sum_{k=1}^n K_{h_3}(x - X_k)}\right]^2\\ &\qquad+ 2E\left[\left(\frac{\sum_{k=1}^n K_{h_3}(x - X_k)(\hat{Z}_k^* - Z_k^*)}{\sum_{k=1}^n K_{h_3}(x - X_k)}\right)\right.\\ &\qquad\left.\times\left(\frac{\sum_{k=1}^n K_{h_3}(x - X_k)(Z_k^* - m(x))}{\sum_{k=1}^n K_{h_3}(x - X_k)}\right)\right]\\ &\quad\equiv A + B + 2C. \qquad (11)\end{aligned}$$

By Lemma 5.2, we have

$$
\begin{aligned}
A &\sim \frac{1}{p^2(x)} E\left[n^{-1} \sum_{k=1}^{n} K_{h_3}(x-X_k)(\hat{f}(X_k, Y_k)-f(X_k, Y_k)) \right]^2 \\
&= \frac{1}{n^2 p^2(x)} \{ nE[K_{h_3}(x-X_1)(\hat{f}(X_1, Y_1)-f(X_1, Y_1))]^2 \\
&\quad + n(n-1)(E[K_{h_3}(x-X_1)(\hat{f}(X_1, Y_1)-f(X_1, Y_1)) K_{h_3}(x-X_2) \\
&\quad \times (\hat{f}(X_2, Y_2)-f(X_2, Y_2))]) \} \\
&\equiv \frac{1}{n^2 p^2(x)} \{ nA_1 + n(n-1) A_2 \}.
\end{aligned}
$$

We first work with A_2. Let $w_{ij} = K_{h_1}(X_i - X_j) K_{h_2}(Y_i - Y_j)(Z_j - f(X_i, Y_i))$, for $i = 1, 2$. Then,

$$
\begin{aligned}
A_2 &\sim \frac{1}{n^2} E\left[\frac{K_{h_3}(x-X_1)}{p(X_1, Y_1)} \left(\sum_{j=1}^{n} w_{1j} \right) \frac{K_{h_3}(x-X_2)}{p(X_2, Y_2)} \left(\sum_{j=1}^{n} w_{2j} \right) \right] \\
&= \frac{1}{n^2} E\left[\frac{K_{h_3}(x-X_1) K_{h_3}(x-X_2)}{p(X_1, Y_1) p(X_2, Y_2)} \right. \\
&\quad \times (2w_{11}w_{21} + (n-2) w_{13}w_{23} + w_{11}w_{22} + w_{12}w_{21} \\
&\quad + 2(n-2) w_{11}w_{23} + 2(n-2) w_{12}w_{23} + (n-2)(n-3) w_{13}w_{24})] \\
&\equiv \frac{1}{n^2} (2A_{21} + (n-2) A_{22} + A_{23} + A_{24} + 2(n-2) A_{25} \\
&\quad + 2(n-2) A_{26} + (n-2)(n-3) A_{27}).
\end{aligned}
$$

Since $w_{11} = (h_1 h_2)^{-1} K^2(0) \varepsilon_1$ and $w_{22} = (h_1 h_2)^{-1} K^2(0) \varepsilon_2$, it is obvious that $A_{23} = A_{25} = 0$. And we have

$$
\begin{aligned}
&E_{Z_1}[\varepsilon_1(Z_1 - f(X_2, Y_2)) \mid X_1, Y_1, X_2, Y_2] \\
&\quad = E[\varepsilon_1^2 \mid X_1, Y_1, X_2, Y_2] = v(X_1, Y_1).
\end{aligned}
$$

Letting $X_1 = x - h_3 z_1$, $X_2 = x - h_1 z_2$, $Y_2 = Y_1 - h_2 z_3$, we can show $A_{21} = O((h_1 h_2 h_3)^{-1})$. Since

$$
\begin{aligned}
&E_{\varepsilon_3}(Z_3 - f(X_1, Y_1))(Z_3 - f(X_2, Y_2)) \\
&\quad = v(X_3, Y_3) + (f(X_3, Y_3) - f(X_1, Y_1))(f(X_3, Y_3) - f(X_2, Y_2)),
\end{aligned}
$$

we have $A_{22} = A_{221} + A_{222}$, where

$$A_{221} = E\left[\frac{K_{h_3}(x - X_1)}{p(X_1, Y_1)} \frac{K_{h_3}(x - X_2)}{p(X_2, Y_2)} v(X_3, Y_3) K_{h_1}(X_1 - X_3) K_{h_2}(Y_1 - Y_3) \right.$$
$$\left. \times K_{h_1}(X_2 - X_3) K_{h_2}(Y_2 - Y_3) \right]$$

and

$$A_{222} = E\left[\frac{K_{h_3}(x - X_1)}{p(X_1, Y_1)} \frac{K_{h_3}(x - X_2)}{p(X_2, Y_2)} K_{h_1}(X_1 - X_3) K_{h_2}(Y_1 - Y_3) \right.$$
$$\times (f(X_3, Y_3) - f(X_1, Y_1)) K_{h_1}(X_2 - X_3) K_{h_2}(Y_1 - Y_3)$$
$$\left. \times (X_3, Y_3) - f(X_2, Y_2)) \right]$$

By making the change of variables $X_1 = X_3 + h_1 z_1$, $X_2 = X_3 + h_1 z_2$, $X_3 = x - h_3 z_3$, $Y_1 = Y_3 - h_2 z_4$ and $Y_2 = Y_3 - h_2 z_5$ and using

$$w(x) = \int v(x, y)\, p(y \mid x)\, dy = \int \frac{v(x, y)\, p(x, y)}{p(x)}\, dy,$$

we find that

$$\begin{aligned} A_{221} &= h_3^{-1} \int K\left(z_3 - \frac{h_1}{h_3} z_1\right) K\left(z_3 - \frac{h_1}{h_3} z_2\right) K(z_1) K(z_2) \\ &\quad \times K(z_4) K(z_5) v(x - h_3 z_3, y_3) p(x - h_3 z_3, y_3) \\ &\quad \times p(x - h_3 z_3, y_3)\, dz_1\, dz_2\, dz_3\, dz_4\, dz_5\, dy_3 \\ &= h_3^{-1} \int K\left(z_3 + \frac{h_1}{h_3} z_1\right) K\left(z_3 + \frac{h_1}{h_3} z_2\right) K(z_1) K(z_2) \\ &\quad \times w(x - h_3 z_3)\, p(x - h_3 z_3)\, dz_1\, dz_2\, dz_3 \\ &= h_3^{-1} \int K\left(z_3 + \frac{h_1}{h_3} z_1\right) K\left(z_3 + \frac{h_1}{h_3}\right) K(z_1) K(z_2) \\ &\quad \times (w(x)\, p(x) + o_p(1))\, dz_1\, dz_2\, dz_3 \\ &= h_3^{-1} k_1 w(x)\, p(x)\, d(h_1/h_3) + o(h_3^{-1}). \end{aligned}$$

By a similar argument as used for A_{221} and from Lemma 5.4, we can show that $A_{222} = o(h_3^{-1})$. Hence $A_{22} = A_{221} + A_{222} = h_3^{-1} k_1 w(x)\, p(x)\, d(h_1/h_3) + o(h_3^{-1})$.

Lastly, we substitute $X_3 = X_1 - h_1 z_1$ and $Y_3 = Y_1 - h_1 z_2$ and use Lemma 5.5 to get

$$\begin{aligned}
A_{27} &= \left\{ E\left[\frac{K_{h_3}(x - X_1)}{p(X_1, Y_1)} \right.\right. \\
&\quad \left.\left. \times K_{h_1}(X_1 - X_3)\, K_{h_2}(Y_1 - Y_3)(f(X_3, Y_3) - f(X_1, Y_1))\} \right] \right\}^2 \\
&= \left\{ E_{X_1, Y_1}\left[\frac{K_{h_3}(x - X_1)}{p(X_1, Y_1)} k_2 \left(\frac{1}{2} h_1^2 t_1(X_1, Y_1) + \frac{1}{2} h_2^2 t_2(X_1, Y_1) \right.\right.\right. \\
&\quad \left.\left.\left. + o(h_1^2 + h_1 h_2 + h_2^2) \right) \right] \right\}^2 \\
&= \frac{1}{4} k_2^2 \left\{ E_{X_1}\left[\frac{K_{h_3}(x - X_1)}{p(X_1)} (h_1^2 s_2(X_1) + h_2^2 g_4(X_1) \right.\right. \\
&\quad \left.\left. + o(h_1^2 + h_2^2 + h_1 h_2)) \right] \right\}^2 \\
&= \frac{1}{4} h_1^4 k_2^2 s_2^2(x) + O(h_1^2 h_2^2 + h_2^4) + o(h_1^4),
\end{aligned}$$

and combining all our calculations,

$$\begin{aligned}
A_2 &= \frac{1}{n h_3} k_1 w(x)\, p(x)\, d(h_1/h_3) + \frac{1}{4} h_1^4 k_2^2 s_2^2(x) \\
&\quad + O(h_1^2 h_2^2 + h_2^4) + o\left(\frac{1}{n h_3} \right) + o(h_1^4).
\end{aligned}$$

Similarly, we can show, under the conditions of the theorem, $A_1 = o(h_3^{-1})$. Hence,

$$A = \frac{1}{n h_3} k_1 \frac{w(x)}{p(x)} d(h_1/h_3) + \frac{1}{4} h_1^4 k_2^2 \frac{s_2^2(x)}{p^2(x)} + O(h_1^2 h_2^2 + h_2^4) + o\left(\frac{1}{n h_3} \right) + o(h_1^4).$$

For B in (11), it is well known (e.g. Collomb [3], Härdle [6]) that

$$B = \frac{1}{n h_3} k_1 \frac{u(x)}{p(x)} + \frac{1}{4} h_3^4 k_2^2 \frac{s_1^2(x)}{p^2(x)} + o\left(\frac{1}{n h_3} \right) + o(h_3^4).$$

Similarly, we observe

$$C = \frac{1}{4} h_1^2 h_3^2 k_2^2 \frac{s_1(x)\, s_2(x)}{p^2(x)} + O(h_2^2 h_3^2) + o\left(\frac{1}{n h_3} \right) + o(h_1^2 h_3^2).$$

Putting everything back to (11), we have

$$\begin{aligned}E[\hat{m}(x)-m(x)]^2 &= \frac{k_1}{nh_3}\frac{w(x)}{p(x)}d(h_1/h_3)+\frac{1}{4}h_1^4k_2^2\frac{s_2^2(x)}{p^2(x)}+\frac{k_1}{nh_3}\frac{u(x)}{p(x)}\\ &\quad+\frac{1}{4}h_3^4k_2^2\frac{s_1^2(x)}{p^2(x)}+\frac{1}{2}h_1^2h_3^2k_2^2\frac{s_1(x)\,s_2(x)}{p^2(x)}\\ &\quad+O(h_1^2h_2^2+h_2^4)+o\left(\frac{1}{nh_3}\right)+o(h_1^4+h_3^4)\\ &=\frac{1}{nh_3}\frac{k_1(u(x)+w(x)\,d(h_1/h_3))}{p(x)}\\ &\quad+\frac{1}{4}h_3^4\frac{k_2^2(s_1(x)+s_2(x)\,h_1^2/h_3^2)^2}{p^2(x)}\\ &\quad+O(h_2^4+h_2^2h_3^2)+o\left(\frac{1}{nh_3}+h_1^4+h_3^4\right).\end{aligned}$$

Equation (5) follows by setting $h_1=\theta h_3$ and $h_2=o(h_3)$.

Proof of Theorem 2.2. From Theorem 2.1, it is easy to see that

(i) If $s_1(x)\,s_2(x)\geqslant 0$, the asymptotic optimal choice of the bandwidth of estimator (1) is $h_3(x)=(D_1(x,\theta^*)/D_2(x,\theta^*))^{1/5}\,n^{-1/5}$ where θ^* minimizes (6) with respect to θ, and the corresponding asymptotic mean squared error is

$$E(\hat{m}(x)-m(x))^2=1.25D_1(x,\theta^*)^{4/5}\,D_2(x,\theta^*)^{1/5}\,n^{-4/5}+o(n^{-4/5}).$$

(ii) If $s_1(x)\,s_2(x)<0$, then if $h_3=o(n^{1/5})$ and $nh_3\to\infty$ and $\theta^2=-s_1(x)/s_2(x)$, the asymptotic mean square error of estimator (1) is $E(\hat{m}(x)-m(x))^2=o(n^{-4/5})$.

And, it is well known (e.g. Collomb [3], Stone [14], Härdle [6]) that, for $h\to 0$ and $nh\to\infty$, the mean squared error of the N-W estimator (2) is

$$E(\tilde{m}(x)-m(x))^2=\frac{D_3(x)}{nh}+\frac{1}{4}D_4(x)\,h^4+o\left(\frac{1}{nh}+h^4\right),$$

where $D_3(x)=k_1\sigma^2(x)/p(x)=k_1(u(x)+w(x))/p(x)$ and $D_4(x)=k_2^2s_1^2(x)/p^2(x)$. The optimal bandwidth h is $(D_3(x)/D_4(x))^{1/5}\,n^{-1/5}$, corresponding to mean squared error

$$E(\tilde{m}(x)-m(x))^2=1.25D_3(x)^{4/5}\,D_4(x)^{1/5}\,n^{-4/5}+o(n^{-4/5}).$$

Theorem 2.2 follows immediately.

The proofs of Theorems 2.3 and 2.4 follow the result of Theorem 2.1.

ACKNOWLEDGMENT

This research is supported in part by the National Science Foundation under Grant DMS-9301193. The author thanks an associate editor and two anonymous referees for their helpful comments.

REFERENCES

[1] Chen, R. (1995). A nonparametric multi-step prediction estimator in Markov structures, *Statistica Sinica*, in press.
[2] Cleveland, W. S., and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting, *J. Amer. Statist. Soc.* **83** 596–610.
[3] Collomb, G. (1977). Quelques propriétés de la méthode du noyau pour l'estimation non-paramétrique de la régression en un point fixé. *Compt. Rendus Acad. Sci. Paris Ser. A* **285** 289–292.
[4] Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. Dekker, New York.
[5] Fan, J. (1992). Design-adaptive nonparametric regression, *J. Amer. Statist. Soc.* **87** 998–1004.
[6] Härdle, W. (1989). *Applied Nonparametric Regression*. Cambridge Univ. Press, Cambridge.
[7] Härdle, W., and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13** 1465–1481.
[8] Hall, P. (1984). Integrated square error properties of kernel estimator of regression functions. *Ann. Statist.* **12** 241–260.
[9] Nadaraya, E. A. (1964). On estimating regression. *Theoret. Probab. Appl.* **9** 141–142.
[10] Parzen, E. (1962). On estimation of a probability density and mode. *Ann. Math. Statist.* **35** 1065–1076.
[11] Scott, D. W. (1992). *Multivariate Density Estimation*. Wiley, New York.
[12] Silvermann, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.
[13] Stone, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–620.
[14] Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.
[15] Watson, G. S. (1964). Smooth regression analysis. *Sankhyā A* **26** 359–372.